Metabolic quirks and the colourful history of the *Euglena gracilis* secondary plastid

Anna M. G. Novák Vanclova1, Martin Zoltner1,2, Steven Kelly3, Petr Soukal1, Kristina Záhonová1,4,5, Zoltán Füissy5, ThankGod E. Ebenezer2,6, Eva Lacová Dobáková5, Marek Eliáš4, Julius Lukes5,7, Mark C. Field2,5 and Vladimír Hampil1

1Faculty of Science, Charles University, BIOCEV, Vestec 252 50, Czechia; 2School of Life Sciences, University of Dundee, Dundee, DD1 5EH, UK; 3Department of Plant Sciences, University of Oxford, Oxford, OX1 3RB, UK; 4Faculty of Science, University of Ostrava, Ostrava 710 00, Czechia; 5Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice 370 05, Czechia; 6School of Biochemistry, University of Cambridge, Cambridge, CB2 1QW, UK; 7Faculty of Science, University of South Bohemia, České Budějovice 370 05, Czechia

**Summary**

- *Euglena* spp. are phototrophic flagellates with considerable ecological presence and impact. *Euglena gracilis* harbours secondary green plastids, but an incompletely characterised proteome precludes accurate understanding of both plastid function and evolutionary history.
- Using subcellular fractionation, an improved sequence database and MS we determined the composition, evolutionary relationships and hence predicted functions of the *E. gracilis* plastid proteome.
- We confidently identified 1345 distinct plastid protein groups and found that at least 100 proteins represent horizontal acquisitions from organisms other than green algae or prokaryotes. Metabolic reconstruction confirmed previously studied/predicted enzymes/pathways and provided evidence for multiple unusual features, including uncoupling of carotenoid and phytol metabolism, a limited role in amino acid metabolism, and dual sets of the SUF pathway for FeS cluster assembly, one of which was acquired by lateral gene transfer from Chlamydiae. Plastid paralogues of trafficking-associated proteins potentially mediating fusion of transport vesicles with the outermost plastid membrane were identified, together with derlin-related proteins, potential translocases across the middle membrane, and an extremely simplified TIC complex.
- The *Euglena* plastid, as the product of many genomes, combines novel and conserved features of metabolism and transport.

**Introduction**

Euglenids are a diverse group of flagellates belonging to the phylum Euglenozoa and have a significant role in the biosphere, as well as being an important model organism and of biotechnological value (Leander et al., 2017). Complex nutritional strategies and an ability to adapt to various environments are major features of the euglenids (Leander et al., 2001; Leander, 2004). The biology of heterotrophic euglenids remains relatively unexplored, but the ease of collection and cultivation of photosynthetic members has made them one of the most widely studied protist groups, despite an absence of a reliable genetic manipulation system. Recent advances in defining the transcriptome and proteome of *Euglena gracilis* promises to improve understanding of the molecular mechanisms employed by this organism and its relatives (Ebenezer et al., 2019).

Euglenophytes harbour green, triple membrane-bound plastids that evolved c. 500 million years ago from an endosymbiont related to Pyramimonadales (Turlen et al., 2009; Jackson et al., 2018). However, it is possible that other symbionts shared both environment and genes with the ancestors of euglenophytes, as a significant proportion of euglenophyte genes are potentially derived from other algal groups (Maruyama et al., 2011; Markunas & Triemer, 2016; Lakey & Triemer, 2017; Ponce-Toledo et al., 2018, Ebenezer et al., 2019), possibly originating during multiple rounds of endosymbiotic gene transfer (Larkum et al., 2007). Higher-order endosymbiotic relationships are the result of complex events that can improve our understanding of organellogenesis. Transfer of genes from endosymbiont to host and establishment of new routes for targeting proteins to the organelle are essential for endosymbiont-to-organelle transformation. Protein targeting systems of secondary plastids are notably similar in otherwise unrelated lineages (Durnford & Gray, 2006; Sommer et al., 2007; Hempel et al., 2009; Spork et al., 2009; Minge et al., 2010; Felsner et al., 2011; Lau et al., 2016), suggesting general mechanistic constraints rather than common descent. In brief, components of the secretory pathway have been recruited for protein translocation across the
outermost plastid membrane: the process starts as co-translational and signal peptide-dependent import into the endoplasmic reticulum (ER) and continues via vesicular transport, either directly or through the Golgi complex (Tonkin et al., 2006; Stiller et al., 2014; Maier et al., 2015), while the major TIC/TOC complex translocases of primary plastids retain their functions at the inner two membranes (van Dooren & Striepen, 2013; Sheiner & Striepen, 2013; Archibald, 2015; Gould et al., 2015; Maier et al., 2015; Böltel & Soll, 2016). In plastids of ‘chromalveolates’, a duplicated version of the ER-associated protein degradation (ERAD) pathway mediates transport across the second-outermost membranes (Spork et al., 2009; Stork et al., 2012; Maier et al., 2015).

In euglenophytes, vesicular transport pathways between the ER, Golgi and plastids are present (Sulli et al., 1999) and have been reconstituted in vitro and biochemically suggest the involvement of GTPases but not SNARE proteins (Sláviková et al., 2005), although the molecular machinery is uncharacterised. A plant-like plastid targeting signal (transit peptide) is essential for plastid import, implying the presence of a TIC/TOC-like pathway (Sláviková et al., 2005). However, in silico analysis of the E. gracilis and Euglena longa transcriptomes identified few TIC subunit homologues, with all TOC subunits failing to be identified even by a sensitive HMMER-based approach using multiple strategies for profile building (Záhonová et al., 2018; Ebenexer et al., 2019), implying a highly divergent or alternative translocation mechanism.

The ultrastructure, metabolism, plastid and mitochondrial genomes of E. gracilis have been studied in great detail (Tessier et al., 1991; Hallick et al., 1993; Muchhal & Schwartzbach, 1994; Jenkins et al., 1995; Doetsch et al., 2001; Geimer et al., 2009; Matešíková-Kováčová et al., 2012; Kuo et al., 2013; Dobáková et al., 2015; Watanabe et al., 2017; Gumińska et al., 2018). Moreover, three transcriptome datasets have been generated recently (O’Neill et al., 2015; Yoshida et al., 2016; Ebenezer et al., 2019), the last coupled with a draft genome and proteomics analysis of whole-cell lysates. Features specific to E. gracilis were uncovered by these studies, which also enabled predictions of plastid protein composition (Záhonová et al., 2018; Ebenexer et al., 2019). However, such predictions rely on both complete understanding of targeting signals, as well as the availability of full-length sequences to encompass N-terminal (and possibly other) plastid targeting peptides. As neither of these criteria is currently met, additional subcellular fractionation evidence is essential for validating predictions and to define a robust proteome. Significantly, the sizes of plastid proteomes vary greatly, from over 1400 proteins in Arabidopsis thaliana to under 400 for the Plasmodium falciparum apicoplast (Huang et al., 2013; Boucher et al., 2018).

Here we isolated E. gracilis plastids and analysed their composition using unbiased MS to identify over 1300 protein groups. We validated many previous predictions, but also uncovered novel metabolic pathways, illuminated targeting mechanisms and identified many lateral gene transfer events that support the sequential endosymbiosis model, or ‘shopping bag hypothesis’, for endosymbiotic evolution.

**Materials and Methods**

**Isolation of plastid fraction**

Plastidal, mitochondrial and peroxisomal fractions of E. gracilis were prepared by gradient ultracentrifugation based on protocols used previously (Davis & Merrett, 1973; Dobáková et al., 2015) and described in detail in Supporting Information Methods S1. The resulting gradient and assessment of fraction quality are documented in Notes S1.2–3.

**MS-based identification and quantification of proteins**

Plastid and mitochondrial fractions were sonicated in NuPAGE LDS sample buffer (Thermo Scientific, Waltham, MA, USA) containing 2 mM dithiothreitol and separated on a NuPAGE Bis-Tris 4–12% gradient polyacrylamide gel (Thermo Scientific) under reducing conditions. Each lane was divided into eight slices that were excised, destained and subjected to tryptic digestion and reductive alkylation. These fractions were subjected to LC-MS/MS on an UltiMate 3000 RSLCnano System (Thermo Scientific) coupled to a Q-exactive mass spectrometer (Thermo Scientific) at the Fingerprints Facility of the University of Dundee. Mass spectra were analysed using MaxQuant (v.1.5, Cox & Mann, 2008), using the predicted translated transcriptome (GEFR00000000.1; Ebenezer et al., 2019). Minimum peptide length was set to six amino acids, isoleucine and leucine were considered indistinguishable and false discovery rates (FDRs) of 0.01 were calculated at the levels of peptides, proteins and modification sites based on the number of hits against a reversed sequence database. Ratios were calculated from label-free quantification (LFQ) intensities using only peptides uniquely mapped to a given protein. If the identified peptide sequence set of one protein overlapped another protein, these two proteins were assigned to the same protein group. P-values were calculated applying t-test-based statistics in Perseus (Cox & Mann, 2008; Tyanova et al., 2016). In total, 3736 distinct protein groups were identified and reduced to 2544 protein groups by rejecting those groups not identified at the peptide level in at least two of three replicates for one organellar fraction. This filtered set includes a cohort of 774 protein groups that were observed in only one organellar fraction (606 in the plastid fraction), and in order to form ratios, a constant small value (0.01) was added to the average LFQ intensities of each fraction (to avoid division by zero).

The resulting CP:MT ratio reflects the enrichment of protein groups in the plastid compared to the mitochondrial fraction and is the main indicator for confidence in plastid localisation of a given protein in this study. For clarity, CP:MT ratios were log10 transformed and values > 3 (CP:MT ratio > 1000) indicated as ‘3+’ in all tables and figures, indicating extremely high or ‘infinite’ enrichment. The purity of the plastid and mitochondrial fractions was assessed based on the distribution and relative abundance of marker proteins (Notes S1, S2). LFQ analysis of the two organellar fractions and whole-cell lysate detected 8216 protein groups and revealed moderate contamination of both mitochondrial and plastid fractions by other cell compartments (Notes...
Proteins enriched in the plastid fraction were sorted into four categories based on log10 CP : MT ratio (0–1, 1–2, 2–3 and 3+), reflecting the robustness/discrimination of plastid targeting and association. All candidates were annotated using BLAST (Altschul et al., 1997) against the NCBI nonredundant protein database (https://www.ncbi.nlm.nih.gov/protein, 08/16 version), and assigned a KO number by KAAS (http://www.genome.jp/tools/kaas/; Moriya et al., 2007). Annotations were validated manually using UniProt (http://www.uniprot.org/) and OrthoFinder (Emms & Kelly, 2015) and corrected as necessary. Proteins with no, very few (fewer than five) or very low-scoring (E-value > 1×10^{-5}) homologues identifiable by BLAST were additionally probed using HHpred against the PDB, COG, ECOD and Pfam databases (https://toolkit.tuebingen.mpg.de/#/tools/hhpred; Soding et al., 2005) and assigned a tentative annotation. Proteins with KO and/or EC numbers assigned, proteins of definite functions in certain metabolic pathways or molecular complexes, as well as proteins with less precise but clearly predicted function were sorted manually into 18 custom-defined categories based on the KEGG pathway classification (Note S3). Thirty-two proteins were discarded from the preliminary dataset of 1377 candidates based on a clearly nonplastid function combined with two proteins were discarded from the preliminary dataset of 1377 based on the KEGG pathway classification (Note S3). Thirty-threeplexes, as well as proteins with less precise but clearly predicted finite functions in certain metabolic pathways or molecular complexes, indicating concordance approaching 90%. By}

N-terminal signal sequence prediction

The hydrophobicity and amino acid composition of the N-termini of plastid-targeted proteins were investigated as described in detail in Methods S1. Analysis was restricted to sequences probably not truncated at the N-terminus, translated in the cytoplasm and imported into the plastid, that is, meeting the following criteria: robust plastid localisation (log10 CP : MT > 1.0) and/or predicted photosynthetic function and encoded by transcripts containing the spliced leader (ATTTTTTTTTCG; Tressier et al., 1991). This cohort consisted of 375 sequences, and a second sequence set of the same size but disregarding plastid localisation was randomly selected from the transcriptome as a control. The differences in amino acid composition were analysed using a χ² test (R Core Team, 2013).

Data availability

All phylogenetic trees calculated are available at: http://www.protiologie.cz/hampplab/data/Novak_Vanclova_trees.zip. The MS proteomics data have been deposited at the ProteomeXchange Consortium via the PRIDE partner repository (Perez-Riverol et al., 2019) with the dataset identifier PXD014767.

Results

A plastid proteome for E. gracilis

The E. gracilis plastid proteome determined here includes 1345 proteins, 48 of which are encoded by the plastid genome (Hallick et al., 1993); 774 (57.5%) could be functionally annotated while 571 (42.5%) have unknown or hypothetical function and 109 of these (8.1%) have no identifiable homologues in the nonredundant database at NCBI (Dataset S1). To assess the completeness of the current dataset we found that 48/67 (72%) plastid encoded proteins were present. We also identified 74/83 (89%) and 113/ 131 (86%) high-confidence plastid-targeted candidates from two earlier studies (Durnford & Gray, 2006; Záhonová et al., 2018; Dataset S2), indicating concordance approaching 90%.
contrast, only 474/1902 (25%) proteins in silico predicted as plastid-targeted from transcriptome data (Ebenezer et al., 2019) were detected by proteomics, and 871 proteins identified by proteomics were absent from this in silico set, suggesting that available prediction tools are inaccurate in the case of Euglena. This could arise partially as a result of N-terminal sequence truncation but could also signify our incomplete understanding of euglenid plastid-targeting signals. It is also of relevance that the plastids analysed here were from light-grown cells only, and previous work has demonstrated that significant changes to protein composition are not accompanied by transcript changes: some proteins are therefore probably absent from the present plastid material (Ebenezer et al., 2019). It is also possible that some plastidial proteins are expressed but their physiological abundance or number of ionisable peptides they produce is too low.

Significantly, this new proteome is of similar size to the A. thaliana plastid proteome (1462 protein groups; Huang et al., 2013), which has itself been determined using a combination of in silico prediction and proteomic analysis, but is substantially larger than proteomes of plastid or plastid-derived organelles from several unicellular organisms, including Chlamydomonas reinhardtii (996; Terashima et al., 2011), Bigelowiella natans (324; Hopkins et al., 2012) and the secondarily nonphotosynthetic parasite Plasmodium falciparum (345; Boucher et al., 2018); Table 1 provides a comparison.

**Biosynthetic pathways embedded in the Euglena plastid**

The current state of knowledge regarding metabolic pathway localisation in Euglena is largely based on biochemical experiments and in silico prediction (recently reviewed by Inwongwan et al., 2019). In this study, we rely on a comprehensive proteomic dataset to reconstruct metabolic pathways localised in the E. gracilis plastid. The 774 functionally annotated proteins were sorted into 18 categories (Fig. 1), and a map of major metabolic pathways and complexes was reconstructed (see Fig. 2 for overall schematic and Notes S5.1–11 for detailed pathways), and although this includes core conserved pathways, we also found features indicative of considerable novelty. As expected, the photosynthetic apparatus is well represented. It should be noted that the precise identification of light-harvesting proteins is not straightforward as these are encoded and translated as polyproteins and are difficult to distinguish using standard proteomic approaches without a specific focus on these proteins (Koziol & Durnford, 2008). In accordance with previous findings (Wildner & Hauska, 1974), no gene or protein for plastocyanin was identified (white circle in e− transport section of Fig. 2; Note S5.2), so electron transfer between the cytochrome bc1 complex and photosystem I relies solely on two isoforms of cytochrome c6. Interestingly, E. gracilis also encodes a homologue of cytochrome c6A (seqid 17930), previously only known from land plants and green algae (Howe et al., 2006). Cytochrome c6A is thought to function as a redox-sensing regulator within the thylakoid lumen and the structure of its N-terminus in E. gracilis is in accordance with this localisation, but was not detected here probably due to low abundance and difficulty with detection even in plants (Howe et al., 2006). Also of note is the presence of two homologues of the alpha subunit of the F-type ATPase in the proteome, a canonical copy encoded by the plastid genome and a divergent copy (seqid 3018) encoded in the nucleus. Finally, the E. gracilis plastid contains two distinct terminal oxidases (seqids 2887 and 18315), components of the chloro-respiration pathway (Nawrocki et al., 2015). One represents the conventional enzyme (phylogenetically affiliated with dinoflagellates), whereas the other falls among mitochondrial alternative oxidases (Note S6). The latter is apparently not a contaminant as it is significantly enriched in the plastid fraction and has a predicted plastid-targeting signal, while at least three different isoforms of the mitochondrial alternative oxidase-bearing mitochondrial transit peptides are also present in E. gracilis.

The Calvin cycle is fully represented, but a route to glucose metabolism is missing due to the lack of glucose 6-phosphate isomerase. This is probably related to the absence of starch synthesis and a switch to the nonplastid β-1,3-glucan paramylon polysaccharide (Barsanti et al., 2001). The key enzyme for paramylon synthesis, glucan synthase-like 2 (GSL2; Tanaka et al., 2017), is absent from the plastid proteome, but its paralogue, GSL1 (seqid 4050), was identified with high confidence. Unfortunately, the substrate specificity and function of GSL1 are unknown. The nearly complete representation of the C5 pathway of tetrapyrrole synthesis, using glutamate as the precursor of the key intermediate aminoluvinate, is in accordance with previous evidence (Gomez-Silva et al., 1985; Koény & Oborník, 2011).

The E. gracilis plastid is also a site for biosynthesis of fatty acids and glycerolipids, the latter including the major plastid phospholipid phosphatidylglycerol and three glycolipids, monogalactosyl-

<table>
<thead>
<tr>
<th></th>
<th>Euglena gracilis (1345)</th>
<th>Arabidopsis thaliana (1462)</th>
<th>Chlamydomonas reinhardtii (996)</th>
<th>Bigelowiella natans (324)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photosynthesis</td>
<td>6.4</td>
<td>9.1</td>
<td>11.9</td>
<td>14.8</td>
</tr>
<tr>
<td>Lipid metabolism</td>
<td>2.9</td>
<td>4.7</td>
<td>2.5</td>
<td>2.2</td>
</tr>
<tr>
<td>Amino acid metabolism</td>
<td>0.7</td>
<td>5.6</td>
<td>5.5</td>
<td>2.2</td>
</tr>
<tr>
<td>Secondary metabolism</td>
<td>2.4</td>
<td>2.9</td>
<td>2.7</td>
<td>2.8</td>
</tr>
<tr>
<td>Tetrapyrrole, cofactor and vitamin metabolism</td>
<td>3.5</td>
<td>5.3</td>
<td>3.2</td>
<td>3.4</td>
</tr>
<tr>
<td>Signalling</td>
<td>3.1</td>
<td>2.3</td>
<td>2.6</td>
<td>3.7</td>
</tr>
</tbody>
</table>

**Table 1** Comparison of the proportions (%) of proteins of selected functional categories in the four plastid proteomes; the number in parentheses under the respective organism name is the total number of identified plastid proteins.
and digalactosyl-diacylglycerol (MGDG and DGDG) and sulfo-
quinovosyl-diacylglycerol (SQDG) (Matson & Meifei 1970; Blea
& Schantz, 1978; Shibata et al., 2018). We reconstructed path-
ways for the synthesis of these compounds, albeit with several
enzymes only predicted by the transcriptome. Interestingly, our
data suggest that *E. gracilis* lacks a key enzyme for SQDG synthe-
sis, UDP-sulfoquinovose synthase (SQD1), which catalyses for-
mation of UDP-sulfoquinovose from UDP-glucose. However,
sulphite, one of the SQD1 substrates, is readily incorporated into
SQDG when incubated with isolated *E. gracilis* plastids (Saidha
& Schiff, 1989), suggesting the presence of an alternative path-
way to UDP-sulfoquinovose, a substrate of the conventional sul-
foquinovosyltransferase (SQD2) catalysing the ultimate step of
SQDG formation (Fig. 2).

Most algal and plant plastids support production of various
amino acids, some of which (such as phenylalanine, tyrosine and
tryptophan) are produced exclusively in the organelle, but the
*Euglena* plastid lacks these pathways and its contribution to
amino acid metabolism is very low. Only an incomplete serine
synthesis pathway was reconstructed, with no identifiable plastid-
targeted phosphoserine aminotransferase. In addition, only phos-
phoserine phosphatase was identified in the proteome, whereas
the plastid localisation of the enzyme catalysing the initial step of
the pathway (phosphoglycerate dehydrogenase, divided into two
contigs, seqids 23072 and 17279) relies on *in silico* prediction.
The *E. gracilis* plastid proteome does contain an isoform of cys-

teine synthase A, but a plastidial version of the enzyme that con-
verts serine to O-acetylserine, the substrate of cysteine synthase A,
was not detected, suggesting that O-acetylserine needs to be

imported into the organelle. The plastid further harbours serine/
glycine hydroxymethyltransferase (SHMT), whose primary role
is probably to provide a formyl group for synthesis of formylme-
thionyl-tRNA for translation initiation in the plastid rather than
to make glycine from serine (Note S5.1). A direct BLAST search
of the whole transcriptome confirmed that the enzymes of amino
acid biosynthesis are indeed present (*E. gracilis* does not require
supplementation by any amino acid when grown autotrophically
on minimal medium; Hall & Schoenborn, 1939), but their local-
isation is nonplastidial, in agreement with the previous *in silico*
prediction (Ebenezer et al., 2019). Possible exceptions are shik-
imate kinase (seqid 11810) (Inwongwan et al., 2019), which
exhibits the typical plastidial targeting signal, and an enzyme in
the shikimate pathway, 5-enolpyruvylshikimate-3-phosphate syn-
thase, which has been previously shown to localise in the
*E. gracilis* plastid (Reinbothe et al., 1994). However, the former
was not detected in the plastid fraction, and the latter is absent from
the transcriptome. Regardless, synthesis of aromatic amino acids
in *E. gracilis* may be fully secured by a complete shikimate path-
way in the cytosol.

In eukaryotes, two pathways are responsible for synthesis of
isopentenyl pyrophosphate (IPP), a precursor to steroids and ter-
penoids: the MVA pathway, localised to the mitochondria/cy-

tosol, and the MEP (DOXP) pathway, compartmentalised into
the plastid (Zhao et al., 2013). *Euglena gracilis* harbours both
pathways, and there is experimental evidence for a role for the
MEP pathway in carotenoid synthesis (Kim et al., 2004). Our
data demonstrate a plastid localisation for the MEP pathway, and
also identify the enzyme for the synthesis of the carotenoid
precursor geranylgeranyl pyrophosphate (GGPP) from IPP and most enzymes required for synthesis of carotenoids previously reported in *Euglena* (Krinsky & Goldsmith, 1960): antheraxanthin, neoxanthin, β-, γ- and ζ-carotene, retinol, lutein and cryptoxanthin. Our proteomic and transcriptomic data, however, failed to provide evidence for β-carotene ketolase synthesising two carotenoids, echinenone and canthaxanthin, that have been detected (Krinsky & Goldsmith, 1960), although some of these enzymes are represented in the transcriptome (marked as grey circles in Fig. 2).

Furthermore, the proteome suggests that IPP, made by the MEP pathway, is used for the synthesis of plastoquinone (via solanesyl-PP), as is usual for plastids in general (Lohr et al., 2012). By contrast, uniquely among plastid-bearing eukaryotes, the MEP pathway of *Euglena* does not provide GGPP for synthesis of phytyl-PP, a precursor of tocopherols (vitamin E), phylloquinone (vitamin K) and Chl (Kim et al., 2004). Our analyses support this earlier experimental finding and provide insight into the alternative pathway. Phytyl-PP is conventionally synthesised by ChlP, a GGPP reductase conserved between eukaryotic and prokaryotic phototrophic organisms (Heyes & Neil Hunter, 2009). The *E. gracilis* plastid proteome and transcriptome lack this enzyme, explaining why GGPP synthesised by the MEP pathway cannot serve as a phytyl-PP precursor in this organism.

![Image](Fig. 2) Overview of *Euglena gracilis* plastid metabolism as reconstructed from the MS-based proteome. Enzymes present in the plastid proteome in at least one isofrom are marked as green circles; grey circles represent enzymes which were identified at the RNA or DNA level (in this study or previously) but are absent from the proteome; white circles represent genes completely absent in *Euglena*; circles marked by the letter P represent genes coded in the plastid genome; the remaining circles represent genes coded in the nucleus. Circles with coloured dots in the middle represent genes with at least one of their isoforms presumably gained via lateral transfer from a donor group other than green algae and ‘miscellaneous’ algae: pale green for chlorarachniophytes, brown for ochrophytes, dark orange for haptophytes, pale orange for cryptophytes, red for rhodophytes, and a combination of the former for proteins of ‘mixed red’ origin (see Supporting Information Notes S4 for a larger and colourblind-friendly version). Multiple overlapping circles represent enzymes composed of multiple subunits with different characteristics. Note that a single protein may be represented by multiple circles due to its role in multiple pathways or reactions. A more detailed schematic including additional pathways and transcript identifiers for each enzyme is available as Notes S5.1–11.

An alternative route to phytyl-PP operates in higher plant chloroplasts, and this recycles phytol released during Chl degradation and significantly contributes to tocopherol biosynthesis (Vom Dorp et al., 2015). This pathway consists of VTE5 (phytol kinase, seqid 13197) and VTE6 (phytyl phosphate kinase, seqid 14381), both of which are present in the plastid proteome. The proteome also contains a homologue of a recently characterised Chl dephytylase (CLD1, seqid 12254; Note S5.5; Lin et al., 2016), which further suggests that phytol is salvaged from degraded Chl. This obviously cannot be the only means of producing phytyl-PP, as the compound itself is required for Chl synthesis. Interestingly, the nonphotosynthetic Chl-lacking *E. longa* retains both VTE5 and VTE6 (GenBank accession numbers GGOE01004182.1 and GGOE01053790.1), suggesting that phytol is a genuine intermediate for de novo phytyl-PP in *Euglena*. As direct phytol synthesis has not been characterised in any organism, how and in which cellular compartment is phytol produced in *Euglena* is thus unclear.

*Euglena gracilis* can produce tocopherol (vitamin E) (Watanabe et al., 2017), and homologues of several enzymes for tocopherol synthesis have been described in the transcriptome (O’Neill et al., 2015). There is some uncertainty concerning the subcellular localisation of *E. gracilis* tocopherol production (Watanabe et al., 2017), which we now resolve by reconstructing...
the full pathway, from the precursors homogentisate and phytyle-
PP to α-tocopherol, by identification of all four enzymes in the
plastid (Fig. 2). Another terpenoid-quinone produced by
E. gracilis is the phylloquinone derivative 5′-monohydroxyphyll-
quinone (Ziegler et al., 1989). In plants and green algae most
steps of phylloquinone synthesis occur in the plastid, except three
reactions from α-succinylbenzoate to dihydroxynaphthoate,
which are catalysed by peroxisomal enzymes (Emonds-Alt et al.,
2017; Cenci et al., 2018). Euglena gracilis encodes a homologue
of the multifunctional protein PHYLLO catalysing all four steps
of the first part of the pathway (seqid 121), but the protein is
absent from the plastid proteome and lacks a targeting prese-
quence, consistent with biochemical evidence placing synthesis
of α-succinylbenzoate in the cytosol (Seeger & Bentley, 1991). The
same study tentatively associated the α-succinylbenzoate to dihy-
droxynaphthoate part of the pathway to the plastid envelope
rather than peroxisomes, but the respective enzymes must be
unrelated or extremely divergent to the canonical peroxisomal
enzymes MenE, MenB and MenI, as their homologues were not
detected in either the plastid proteome and the transcriptome.
Genome analyses indicate the existence of a novel pathway of
dihydroxynaphthoate synthesis in some algae (e.g. glaucophytes)
and cyanobacteria (Cenci et al., 2018), so it is possible that E.
gracilis converts α-succinylbenzoate to dihydroxynaphthoate by
employing enzymes of this route. The final part of the pathway,
starting with phytlation of dihydroxynaphthoate, is predicted as
plastid-localised in E. gracilis (Note S5.5), although only one of
the respective enzymes (MenA, seqid 7550) was identified in the
proteome and the identity of the enzyme catalysing the presum-
ably final step (phyloquinone hydroxylation) is unknown.

Protein import and vesicle targeting to the Euglena plastid
A major plastid import pathway is mediated by the TIC/TOC
translocation complexes. Surprisingly, only two subunits were
identified in the transcriptome: Tic32 and three paralogues of
Tic21 (Záhonová et al., 2018). Proteomics confirmed the plastid
localisation for all Tic21 isoforms, but not Tic32 (Fig. 3). Tic21 is
structurally similar to Tic20, the main channel-forming sub-
unit in canonical TIC, and it was sporadically isolated in complex
with Tic20 and other central subunits. This suggests that Tic21
is nonessential in plant plastids (Teng et al., 2006; Kikuchi et al.,
2009; Nakai, 2018), but it assuming a main translocase function
in a highly reduced complex is still conceivable. Tic32, by con-
trast, is a regulatory subunit with versatile enzymatic activity
(Balsera et al., 2010) that could readily serve a TIC-independent
purpose or was simply not detected.

Two proteins (seqids 13308 and 11030), representing a novel
subgroup of rhomboid-related pseudoproteases distantly related
to derlins of the ERAD pathway, were also found (Notes S11.1,
S14). In ‘chromalveolate’ plastids, the ERAD machinery is par-
ially duplicated and recruited for protein transport forming a
system termed SELMA, with two derlin family proteins (DerL-1
and DerL-2) presumably constituting the protein-conducting
channel (Maier et al., 2015). The E. gracilis proteins, however,
represent distinct and more distantly related homologues of
ERAD derlins. Although the phylogeny is poorly supported, it
does suggest that these proteins originated by duplication of a
host gene, because a third parologue (seqid 22665, absent from
the plastid proteome) is present in the transcriptomes of
euglenids including the heterotroph Rhodomonas costata (Notes
S11.2, S14). It is intriguing to speculate that the detected derlin-

![Fig. 3 Reconstruction of the Euglena gracilis plastid protein import
machinery: the N-terminal signal-peptide (SP, red) bearing nuclear-
encoded plastid-targeted protein (blue chain) is synthesized on the rough
endoplasmic reticulum (ER). If the N-terminus does not contain a stop
transfer signal (STS), it is co-translationally imported into the ER lumen
as indicated. Otherwise, the major part of the protein remains in the cytosol
while being anchored in the ER membrane. The SP is cleaved by signal
peptidase in the ER lumen. The protein then passes through the Golgi and
is loaded into/onto a vesicle which eventually fuses with the outermost
plastid membrane. Plastidial homologues of GOSR and Rab5 GTPase
might mediate the fusion. The protein passes the middle membrane via an
unknown mechanism dependent on transit peptide (TP, purple) but not
TOC complex, possibly employing derlin-like proteins (DerL). It then
passes the inner membrane via highly reduced TIC translocase, possibly
consisting of multiple isoforms of a single subunit. The protein is folded
and has its transit peptide cleaved in the plastid stroma and, in case
additional signal is revealed, enters the thylakoid membrane or lumen via
the TAT, SEC or SRP/Alb3 pathway. Each circle represents a putative
subunit described in model plastids: green circles represent proteins
identified in the plastid proteome, different colour shades code the
CP: MT ratio, the darkest shade depicting the most credible evidence for
plastid localisation, and grey circles represent subunits with transcript-level
evidence only. Finally, white circles represent subunits which are
completely absent.
like proteins in the *E. gracilis* plastid gained a role similar to SELMA through convergent evolution. The delivery of some plastid-targeted proteins in euglenophytes involves fusion of Golgi-derived vesicles with the outermost plastid membrane and requires GTP hydrolysis, but is not *N*-ethylmaleimide-sensitive, suggesting the involvement of a Rab/ARF GTPase but not a requirement for conventional SNARE-mediated fusion (Sláviková *et al.*, 2005). Multiple paralogues of some membrane trafficking components are present in *E. gracilis*, and we previously speculated that some of these may be involved in plastid transport (Ebenexer *et al.*, 2019). Interestingly, the plastid proteome contains homologues of coat complex subunits, several Rab GTPases and SNARE proteins, some of which cannot be dismissed as contaminants. For example, SNARE protein GOSR1 has two paralogues in *E. gracilis*, one of which is detected in the plastid proteome with log$_{10}$ CP : MT ratio > 3 (seqid 18194), and the other (seqid 19152) was not detected; thus the apparent insensitivity to NEM might potentially be a technical issue or suggest that GOSR1 functions in an NSF-independent manner. A second notable candidate is Rab5, a conserved Rab GTPase associated with the endosomal system (Langemeyer *et al.*, 2018), again with convincing log$_{10}$ CP : MT ratio > 3. These two proteins thus may play roles in targeting protein-transporting vesicles to the outermost plastid membrane (Fig. 3), although the current state of understanding of this pathway is rudimentary.

**Evolutionary origins of plastid proteins**

The semi-automatic phylogenetic pipeline recovered 379 (28%) potential lateral gene transfer (LGT) candidates: 346 of these (91%) are related to phototrophic eukaryotes (primary or secondary algae), 20 (5%) to other eukaryotes and 13 (3%) to prokaryotes (Fig. 4). Most algae-affiliated proteins are related to green algae (95, 28%), as expected given the plastid ancestry. The number of proteins likely to be derived from rhodophytes, cryptophytes and glaucophytes is very low (around 1–2% each), arguably at the level of technical error. By contrast, protein cohorts related to chlorarachniophytes, ochrophytes and haptophytes (6–7% each, 19% in total) are sufficient to be considered as genuine contributions to the *E. gracilis* plastid. Some proteins were related to a clan composed of several groups of phototrophs, and therefore assigned to one of several 'miscellaneous' categories (Fig. 4). The distribution of the metabolic functions of proteins affiliated to nongreen algae has no clear pattern (Fig. 2). A comparison of these results to the previously published findings on the same topic (Markunas & Triemer, 2016; Ponce-Toledo *et al.*, 2018) is available in Dataset S2.

Additionally, 37 plastidial proteins were determined as affiliated to Discoba, that is pre-existing proteins that might have been repurposed or duplicated and recruited for plastidial function. These include mostly membrane transporters, vesicle-associated proteins and chaperones (see Dataset S1). Most of these have log$_{10}$ CP : MT ratios close to zero and might represent dual-localised proteins or contaminations from other cell fractions, although the functions of the more credible candidates suggest that the host genome contributed to the plastid proteome mostly by housekeeping or regulatory functions through ER/Golgi-like proteins and transporters.

Several of the 13 prokaryote-derived proteins encode significant and unexpected functions. For example, four components (SufS, SufB, SufC and SufD) of the SUF system for iron–sulphur (FeS) cluster assembly were found. In the transcriptomic data, we also recovered the fifth SUF pathway component, SufE, also of prokaryotic origin. Three of these proteins were specifically related to Chlamydiae in which the SufBCDS genes form a single operon that could have been laterally transferred as one unit from a single source. It is unknown whether these genes cluster together as they were too fragmented in the *E. gracilis* genome assembly (Ebenexer *et al.*, 2019). In general, the SUF pathway endosymbiотically derived from cyanobacteria is essential and universally present in plastids, supplying multiple core enzymes with FeS clusters (Lu, 2018). The chlamydiae-like set of proteins represents a second path for FeS cluster assembly alongside the ancestral plastidial cyanobacteria-related SUF system homologues (Fig. 5; Notes S7, S8). This apparent redundancy does not represent a transient state and/or a result of neutral evolution because these two parallel SUF systems are also present in *E. longa* and *Eutreptiella* spp. (Notes S7, S8). The relative protein abundance suggests that chlamydiae-like SUF proteins are generally more abundant than the cyanobacteria-like cohort, indicating that the chlamydiae-like pathway is active and contributes to plastid FeS metabolism (Fig. 5). The co-occurrence of two SUF pathways in euglenophyte plastids may indicate functional diversification, for example with one in the stroma in common with other plastids (Lu, 2018), and the second restricted to one of the intermembrane spaces, serving co-localised FeS proteins. This hypothesis of spatial separation of the two SUF pathways is further supported by the fact that the chlamydiae-like proteins, unlike the other cohort, possess N-terminal extensions which are generally shorter and do not conform to a typical plastid-targeting domain (Note S15). Similarly, a parallel CIA pathway for FeS cluster assembly was described in cryptophytes and proposed to function in the periplastidial compartment (PPC), which contains a nucleomorph and represents the remnant endosymbiont cytoplasm (Grosche *et al.*, 2018).

A further novel aspect of the FeS cluster assembly machinery in *E. gracilis* is the presence of an ABC transporter (seqid 3116) in the plastid proteome, homologous to the mitochondrial Atm1 protein required for the export of unspecified FeS cluster intermediates to the cytosol. The presence of this transporter suggests that such intermediates are transported across a membrane, either into the plastid or between subcompartments (Dataset S1).

While direct experimental data are necessary to understand the function of the second FeS pathway, two additional chlamydial-like proteins were also found, both with orthologues in *E. longa* and *Eutreptiella*. One, an isoform of ferredoxin (divided into two contigs, seqids 37420 and 61063 in our data, seqid 74687 in Yoshida *et al.* (2016); Dataset S2; Note S9), is itself an FeS protein, and may represent a client of the chlamydial-like SUF. The second is the alpha subunit of NADPH-dependent sulphite
reductase (CysJ). Significantly, the beta subunit of this enzyme (CysI) contains an Fe4S4 cluster (Smith & Stroupe, 2012) and seems to be related to spirochaetes (Note S10). The plastid localisation of sulfite reductase is itself notable, as previous biochemical studies associated sulphate assimilation, including the reaction catalysed by sulphite reductase, with the mitochondrion in *E. gracilis* (Saidha *et al*., 1988). The plastid localisation of sulphite reduction is consistent with the presence of cysteine synthase in this organelle, as this enzyme assimilates hydrogen sulphide produced in this reaction.

N-terminal plastid-targeting sequences in *E. gracilis*

We took advantage of the confidently identified plastid-targeted proteins to re-evaluate characteristics of *E. gracilis* plastid-targeting sequences. N-terminal sequences of 375 preproteins with well-supported plastid localisation (log_{10} CP:MT ratio > 1 or photosynthetic function) and nontruncated N-termini, verified by the presence of the spliced leader, were analysed. The presence of one or two hydrophobic domains was confirmed using the Kyte–Doolittle hydrophobicity score (Kyte & Doolittle, 1982)
in positions 9–60 (putative signal peptide, SP) and 97–130 (putative stop-transfer signal, STS) (Fig. 6). The sequences were sorted into previously defined classes (Durnford & Gray, 2006): 47% class I (SP and STS), 37% class II (SP only), and 16% exhibiting neither of these hydrophobic domains, referred to as ‘unclassified’. The proportion of ‘unclassified’ preproteins is relatively high and may suggest a significant cohort of plastid proteins with nonconventional targeting signals. No significant correlation between predicted function or preprotein classes was observed.

The canonical euglenophyte plastid-targeting sequence includes a region directly downstream of the SP that exhibits features of a plastid transit peptide (transit peptide-like region, TPL) (Inagaki et al., 2000; Sláviková et al., 2005; Durnford & Gray, 2006), which is surprising, given the apparent absence of a conventional TOC complex in these organisms. To better understand the role of TPL region we investigated its sequence characteristics more closely. The putative TPL region was selected from the 375 plastid proteins, either based on the actual positions of the hydrophobic domains or from the TPL region (61–95) estimated from the overall hydrophobicity profile of all studied N-termini. Amino acid compositions of these putative TPLs differed significantly from their respective mature protein regions (Fig. 7). This contrasts with the control sets of 375 randomly selected proteins, in which the only significant difference between the putative TPL region and the rest of the protein was a relative enrichment in serine (for full results for both sets and all classes of preproteins, see Notes S12, S13).

Discussion

The E. gracilis plastid proteome described here is similar in size to that of A. thaliana (Huang et al., 2013), but larger than C. reinhardtii (Terashima et al., 2011) and B. natans (Hopkins et al., 2012), suggesting that its functional capacity might be higher than other unicellular phototrophs, which could be connected to the metabolic versatility of euglenophytes. More proteins lacking readily identifiable homologues are present, compared to C. reinhardtii, for example, and the proportion of proteins belonging to equivalent functional categories clearly differs, suggesting divergence in the relative significance of functions. While the former is in part a result of experimental organism bias, it also indicates the importance of analysis of organelles from divergent lineages to gain both an accurate understanding of function in the organism itself as well as across eukaryotic systems.

The euglenid plastid originated from pyramimonadalean green alga (Turmel et al., 2009; Jackson et al., 2018), but a proportion of plastid proteins show distinct phylogenetic affinity. Being noted previously (Maruyama et al., 2011; Markunas and Triemer, 2016; Lakey and Triemer, 2017; Ponce-Toledo et al., 2018), it was suggested that these genes were acquired from ‘chromophyte’ prey or a symbiont by the common ancestor of eukaryovorous and/or phototrophic euglenids. The present set of experimentally determined plastid proteins both supports this hypothesis and allows semiquantitative evaluation of contributions from phototrophic eukaryotes. Cohorts related to chlorarachniophytes, ochrophytes and haptophytes account for over 60 E. gracilis plastid proteins (19% of the algal LGT candidates).

These genes could have originated: from the eukaryovorous ancestor of euglenids, sensu the ‘you are what you eat’ hypothesis (Doolittle, 1998); from initial stages of endosymbiotic integration when the euglenid host was presumably obligatorily mixotrophic (much like Rapaza viridis, Yamaguchi et al., 2012), and such transfers could have compensated for the reductive evolution of the endosymbiont genome, as proposed for the chromatophore of Paulinella (Marin et al., 2005; Nowack et al., 2016); from gene transfers from a cryptic endosymbiont, kleptoplastid or even true plastid putatively present in the euglenophyte ancestor and replaced by the extant organelle, in the ‘shopping bag’ (Larkum

![Fig. 6](image-url) The two hydrophobic domains in the N-termini of 375 well-supported Euglena gracilis plastid-targeted proteins with complete N-terminus as seen on the graph of average hydrophobicity score per position compared to the negative control (375 randomly selected proteins of E. gracilis). The range of the signal peptide (SP) was estimated as positions 9–60 (note that the first and last six positions of the 160 amino acid long sequence are missing as a result of window-based scoring; position 7 is the first position with score assigned); the range of the stop-transfer signal (STS) was estimated as positions 97–130. The hydrophobic peak representing the STS is clearly less accentuated than the one representing SP, reflecting the fact that it is present in just a subset of plastid-targeted proteins termed class II pre-proteins.
et al., 2007) or ‘red carpet’ (Ponce-Toledo et al., 2019) sense and similar to the cases of serial endosymbioses and overall plastid fluidity in dinoflagellates (Saldarriaga et al., 2001; Yoon et al., 2005; Takano et al., 2008; Matsumoto et al., 2011; Xia et al., 2013; Burki et al., 2014; Dorrell & Howe, 2015); or from horizontal transfers from euglenids to the other algal group. The last is a possible model for proteins affiliated to chlorarachniophytes given that euglenids are estimated as >200 million years younger than this group (Jackson et al. 2018).

The proportion of credible plastidial proteins affiliated to Discoba, probably representing ancestral euglenozoan proteins newly recruited for plastidial function, is relatively low. The set is noticeably biased towards proteins related to membrane and vesicular transport, which, given the methodological constrains of this study, can be interpreted either as the result of the role of the host endomembrane system in the transport of biomolecules, or sometimes more obviously as a mere contamination of both analysed fractions by other subcellular compartments, namely ER, Golgi and/or vesicles.

The presence of proteins of prokaryotic affiliation is noteworthy and unexpected. Many point at the Chlamydiae as the donor group, the most notable example being a complete second SUF pathway of FeS cluster assembly and hence potentially the result of a single transfer event. A surprisingly high number of chlamydial-like proteins are present in primary plastids of plants and algae (Moustafa et al., 2008; Becker et al., 2008), which led to the ‘ménage à trois’ hypothesis proposing that a chlamydial endosymbiont was directly implicated in integration of a nascent plastid, and in some manner may have even been essential to the process (Facchinelli et al., 2013; Cenci et al., 2016). Whether this could also be the case for secondary plastid establishment remains an intriguing possibility.

Fig. 7 Statistical comparison of the overall amino acid frequencies in the putative transit peptide region and the putative mature chain of the same protein for the set of 375 Euglena gracilis proteins regardless of their classification for the plastid protein sample in comparison to the negative control (shades of blue and orange, respectively (a)). The same comparison was also performed separately for the plastid proteins of class I, class II and ‘unclassified’ (shades of red, orange and yellow, respectively (b)). The vertical axis represents the normalised $\chi^2$ sum reflecting whether the amino acid frequency is higher or lower than expected (positive or negative values) and the relative degree of its enrichment or depletion. Medium coloured bars represent statistically significant results with $P < 0.01$, while dark coloured bars represent those with $P < 10^{-5}$. 

New Phytologist (2020) 225: 1578–1592
New Phytologist © 2019 New Phytologist Trust

www.newphytologist.com
While the E. gracilis plastid TPL region is not conserved at the sequence level, there is a characteristic amino acid composition (Fig. 7). In addition to previously described positive net charge, enrichment of hydroxy residues and alanine common with the TP of plant plastid proteins (Bruce, 2000; Durnford & Gray, 2006; Patron & Waller, 2007; Felsner et al., 2010; Li & Teng, 2013), there is a paucity of nonpolar residues and the high proline content that may confer distinctive secondary structure (MacArthur & Thornton, 1991). Interestingly, proline residues in the TP were recently proposed to be important for correct import of plastid proteins with multiple transmembrane domains in plants (Lee et al., 2018), but it is unclear what the significance of general proline-rich in TPs, as observed here, could be. The additional properties of the TPL are probably recognised by an as yet unknown receptor, consistent with a major reorganisation to translocation systems, namely a reduced TIC/TOC complex and the presence of derlin-like proteins. The absence of unambiguously identifiable additional components, such as Npl4 and Ubx, leaves it unclear if a translocon analogous to SELMA is functional in E. gracilis. Nevertheless, we believe it is a noteworthy candidate for components of the unknown euglenophyte plastid protein import system, specifically as translocases of the middle plastid membrane (Fig. 3), filling a gap in the protein import machinery caused by the absence of TOC. These findings also hint at the possibility of the middle membrane being derived from the host endomembrane system and not the outer cyanobacterial-like membrane of the ancestral primary plastid. Most secondary plastids keep all four membranes and it is not clear how probable or improbable the loss of one of the cyanobacterial membranes is. However, the recently described structure of the plastid of the stramenopile Chrysoparadoxa suggests that this membrane topology can indeed arise (Wetherbee et al., 2018).

In summary, this is the first report of a proteome of a euglenid plastid based on direct organellar isolation and proteomics. Novel import systems, metabolic pathways and the presence of significant transfers from prokaryotic genomes all contribute towards a complex and divergent organelle which has a major impact on the biosphere.

Acknowledgements

We thank Anna Nenarokova (Biology Centre, České Budějovice) for suggestions and recommendations for data analysis and Dougie Lamont and colleagues at the Proteomics Facility of the University of Dundee for excellent work. This work was supported by Czech Grant Agency (16-25280S to VH, 17-21409S to ME); ERC CZ (award LL1601 to JL); ERD Funds (project OPVVV CZ.02.1.01/0.0/0.0/16_019/0000759 to JL, ME and VH); Yousef Jameel Academic Program (through the Yousef Jameel PhD Scholarship); the Cambridge Commonwealth; European and International Trust; the Cambridge University Student Registry; the Cambridge Philosophical Society (all to TEE) and the Medical Research Council (Grant no. P009018/1 to MCF). Computational resources were supplied by the Ministry of Education, Youth and Sports of the Czech Republic under the Projects CESNET (Project no. LM2015042) and CERIT-Scientific Cloud (Project no. LM2015085) provided within the programme Projects of Large Research, Development and Innovations Infrastructures.

Author contributions

MCF, JL and VH conceived the original research plans; JL and MCF supervised the experiments; ELD performed the cell fractionation; MZ and SK performed MS analysis; AMGNV and TGE performed protein annotation and sorting; AMGNV, KZ, ZF and ME interpreted the annotation results; AMGNV performed the signal domain analysis; PŠ performed the phylogenetic analysis; AMGNV conceived the project and wrote the article with contributions of all the authors; VH, ME, MCF and JL supervised and complemented the writing. VH agreed to serve as the author responsible for contact and ensuring communication.

ORCID

ThankGod E. Ebenezer https://orcid.org/0000-0001-9005-1773
Marek Eliáš https://orcid.org/0000-0003-0666-6542
Mark C. Field https://orcid.org/0000-0002-4866-2885
Zoltán Füssy https://orcid.org/0000-0002-0820-6359
Vladimir Hamplí https://orcid.org/0000-0002-5430-7564
Steven Kelly https://orcid.org/0000-0001-8583-5362
Julius Luket https://orcid.org/0000-0002-0578-6618
Anna M. G. Nová Vanclová https://orcid.org/0000-0002-1613-1625
Petr Soukal https://orcid.org/0000-0002-3125-253X
Kristina Záhonová https://orcid.org/0000-0002-5766-0267
Martin Zoltner https://orcid.org/0000-0002-0214-285X

References


Inagaki J, Fujita Y, Hase T, Yamamoto Y. 2000. Protein translocation within chloroplast is similar in *Euglena* and higher plants. *Biochemical and Biophysical Research Communications* 277: 436–442.


